2

# Segment-based Acoustic Models for Continuous Speech Recognition

## Progress Report: January – March 1993

N00014-92-J-1778

submitted to
Office of Naval Research
and
Defense Advanced Research Projects Administration
5 April 1993

by
Boston University
Boston, Massachusetts 02215

**Principal Investigators**

Dr. Mari Ostendorf
Assistant Professor of ECS Engineering, Boston University
Telephone: (617) 353-5430

Dr. J. Robin Rohlicek
Scientist, BBN Inc.
Telephone: (617) 873-3894

**Administrative Contact**

Maureen Rogers, Awards Manager
Office of Sponsored Programs
Telephone: (617) 353-4365

26pp

# Executive Summary

This research aims to develop new and more accurate acoustic models for speaker-independent continuous speech recognition, by extending previous work in segment-based modeling and by introducing a new hierarchical approach to representing intra-utterance statistical dependencies. These techniques, which are more costly than traditional approaches because of the large search space associated with higher order models, are made feasible through rescoring a set of HMM-generated N-best sentence hypotheses. We expect these different acoustic modeling methods to result in improved recognition performance over that achieved by current systems, which handle only frame-based observations and assume that these observations are independent given an underlying state sequence.

In the third quarter of the project, in coordination with a related DARPA-NSF grant (NSF no. IRI-8902124), we have: further investigated techniques for improving the baseline stochastic segment model (SSM) system, including exploration of several alternatives for tied mixture modeling and development of new faster training techniques, as well as further development of multiple pronunciation word models; and started porting our recognition system to the new Wall Street Journal task, a standard task in the ARPA community.

Though much of our effort has gone towards moving to the new task, we have also achieved a 9% reduction in error on the Resource Management corpus in the last three months for the SSM system. We currently report 3.6% word error on the October 1989 Resource Management test for the SSM alone, and 3.1% word error for the combined SSM-HMM system.

# Contents

Principal Investigator Name: Mari Ostendorf
PI Institution: Boston University
PI Phone Number: 617-353-5430
PI E-mail Address: mo@raven.bu.edu
Grant or Contract Title: Segment-Based Acoustic Models for Continuous Speech Recognition
Grant or Contract Number: ONR-N00014-92-J-1778
Reporting Period: 1 January 1993 – 31 March 1993

# 1 Productivity Measures

- Refereed papers submitted but not yet published: 1
- Refereed papers published: 0
- Unrefereed reports and articles: 2
- Books or parts thereof submitted but not yet published: 0
- Books or parts thereof published: 0
- Patents filed but not yet granted: 0
- Patents granted (include software copyrights): 0
- Invited presentations: 0
- Contributed presentations: 0
- Honors received:
  M. Ostendorf chosen to chair the 1996 ARPA Workshop on Human Language Technology
  M. Ostendorf invited to participate in the DoD workshop on Robust Speech Analysis
- Prizes or awards received: 0
- Promotions obtained: 0
- Graduate students supported $\geq$ 25% of full time: 2
- Post-docs supported $\geq$ 25% of full time: 0
- Minorities supported: 0

Principal Investigator Name: Mari Ostendorf
PI Institution: Boston University
PI Phone Number: 617-353-5430
PI E-mail Address: mo@raven.bu.edu
Grant or Contract Title: Segment-Based Acoustic Models for Continuous Speech Recognition
Grant or Contract Number: ONR-N00014-92-J-1778
Reporting Period: 1 January 1993 – 31 March 1993

# 2 Summary of Technical Progress

## Introduction and Background

In this work, we are interested in the problem of large vocabulary, speaker-independent continuous speech recognition, and specifically in the acoustic modeling component of this problem. In developing acoustic models for speech recognition, we have conflicting goals. On one hand, the models should be robust to inter- and intra-speaker variability, to the use of a different vocabulary in recognition than in training, and to the effects of moderately noisy environments. In order to accomplish this, we need to model gross features and global trends. On the other hand, the models must be sensitive and detailed enough to detect fine acoustic differences between similar words in a large vocabulary task. To answer these opposing demands requires improvements in acoustic modeling at several levels. New signal processing or feature extraction techniques can provide more robust features as well as capture more acoustic detail. Advances in segment-based modeling can be used to take advantage of spectral dynamics and segment-based features in classification. Finally, a new structural context is needed to model the intra-utterance dependence across phonemes.

This project addresses some of these modeling problems, specifically advances in segment-based modeling and development of a new formalism for representing inter-model dependencies. The research strategy includes three thrusts. First, speech recognition is implemented under the N-best rescoring paradigm [1], in which the BBN Byblos system is used to constrain the segment model search space by providing the top N sentence hypotheses. This paradigm facilitates research on the segment model through reducing development costs, and provides a modular framework for technology transfer that has already enabled us to advance state-of-the-art recognition performance through collaboration with BBN. Second, we are working on improved segment modeling at the phoneme level [2, 3, 4] by developing new techniques for robust context modeling with Gaussian distributions, and a new stochastic formalism – classification and explicit segmentation scoring – that more effectively uses segmental features. Lastly, we plan to investigate hierarchical structures for representing the intra-utterance dependency of phonetic models in order to capture speaker-dependent and session-dependent effects within the context of a speaker-independent model.

## Summary of Recent Technical Results

In the first half of Year 1, we focused on improving the performance of the basic segment word recognition system. In brief, the accomplishments of that period included: improvments to the N-Best rescoring technique by introducing score normalization; development of a method for clustering contexts to provide robust context-dependent model parameter estimates; extensions to the classification and segmentation scoring formalism to handle context-dependent models with long-range acoustic features; and extension of the two level segment/microsegment formalism and assessment of trade-offs in mixture vs. trajectory modeling. In addition, we began investigating algorithms for the automatic generation of multiple-pronunciation word networks and the use of tied mixtures in the segment model.

The research efforts during this quarter, again supported in part by a related DARPA-NSF grant (NSF no. IRI-8902124), have focused on furthering the multiple-pronunciation and tied mixture studies, writing up previous work (see attached papers), and mainly on porting our recognition system to the Wall Street Journal (WSJ) domain. In particular, we have:

*Investigated the use of different phone sets and multiple-pronunciation networks:* A facility for generating multiple pronunciations, developed under NSF grant number IRI-8805680 for obtaining high quality phonetic alignments of speech, was extended and reimplemented for recognition applications. Robust pronunciation probabilities were estimated by tying the probabilities of transitions generated by the same rule. Although no improvements were obtained on the RM corpus, we plan to further investigate the use of multiple pronunciations in the WSJ domain. In addition, we investigated various phone sets, finding that it was useful to include separate models for fronted and non-fronted schwa and syllable initial and final /r/ and /l/, but not useful to separately model lexical stress. (The lexical stress result was unexpected, but has been confirmed by researchers at other sites.)

*Investigated the use of tied mixture distributions:* Though many HMM recognition systems now use tied mixture distributions, the trade-offs of various modeling assumptions (e.g. feature correlation) and parameter estimation conditions had not been fully investigated. We therefore explored several of these issues, experimenting in the context of an SSM with frame-level mixtures. We found that full covariance component distributions outperform diagonal covariances, joint modeling of cepstra and differenced cepstra gives slightly better results than treating these features independently, parameter initialization using a sampling of context-dependent models gives better results than K-means initialization, and re-estimation of all parameters gives improved peformance over re-estimation of just the mixture weights and/or distribution means. In addition, we developed two mechanisms for reducing training time: training only detailed context models and computing other levels of context conditioning as marginals of these distributions, and a course-grain parallel implementation of training that scales linearly with the number of workstations available. These results are reported in [5], which is attached. Overall, we achieved a 20% reduction in word error over our baseline SSM results [4] on the Resource Management task. We also extended the tied mixture formalism to handle segment and micro-segment component distributions and are currently experimenting with micro-segment level tied mixtures.

*Ported the SSM word recognition system to the Wall Street Journal task domain:* The effort to port our recognition system to the WSJ domain involved modifying functions to maintain compatibility with BBN, modifying I/O formats to handle the new dictionary for eventually evaluating the multiple-pronunciation networks, and porting both the tied-mixture and non-mixture versions of the SSM trainer and recognizer. The porting activity, which has in part served to train a new student, is largely complete and we expect to have our first recognition results shortly.

Our current best result on the Resource Management task is based on the tied-mixture system, which achieves 3.6% word error on the October 1989 test set (a slight improvement over our best result in December) and 7.3% word error on September 1992 test set. Our best combined HMM-SSM results on RM are the same as reported in December: 3.1% on the October 1989 test set and 6.1% word error on the September 1992 test set.

### Future Goals

Based on the results of the past year and our original goals for the project, we have set the following goals for the remainder of Year 1: (1) continue the effort to move to the 5000-word Wall Street Journal task; (2) investigate the use of tied mixtures at the microsegment level; (3) investigate unsupervised adaptation in the WSJ task domain; and (4) develop the hierarchical model formalism together with methods for robust parameter estimation.

## References

[1] M. Ostendorf, A. Kannan, S. Austin, O. Kimball, R. Schwartz, J. R. Rohlicek, "Integration of Diverse Recognition Methodologies Through Reevaluation of N-Best Sentence Hypotheses," *Proc. of the DARPA Workshop on Speech and Natural Language*, pp. 83-87, February 1991.

[2] M. Ostendorf and S. Roukos, "A Stochastic Segment Model for Phoneme-Based Continuous Speech Recognition," *IEEE Trans. Acoustics Speech and Signal Processing, Dec. 1989.*

[3] S. Roucos, M. Ostendorf, H. Gish, and A. Derr, "Stochastic Segment Modeling Using the Estimate-Maximize Algorithm," *IEEE Int. Conf. Acoust., Speech, Signal Processing*, pages 127–130, New York, New York, April 1988.

[4] M. Ostendorf, A. Kannan, O. Kimball and J. R. Rohlicek, "Continuous Word Recognition Based on the Stochastic Segment Model," *Proceedings of the 1992 DARPA Workshop on Artificial Neural Networks and Continuous Speech Recognition*, September 1992.

[5] "On the Use of Tied Mixture Distributions," O. Kimball and M. Ostendorf, to appear in *Proceedings of the ARPA Workshop on Human Language Technology*, 1993.

Principal Investigator Name: Mari Ostendorf
PI Institution: Boston University
PI Phone Number: 617-353-5430
PI E-mail Address: mo@raven.bu.edu
Grant or Contract Title: Segment-Based Acoustic Models for Continuous Speech Recognition
Grant or Contract Number: ONR-N00014-92-J-1778
Reporting Period: 1 January 1993 – 31 March 1993

## 3 Publications and Presentations

Papers written during the reporting paper include a site report and a conference paper, in association with the March 1993 ARPA Workshop on Human Language Technology, and a correspondence paper that has been recently submitted for publication, as listed below. Copies of these papers are included with the report.

- "Segment-Based Acoustic Models for Continuous Speech Recognition," M. Ostendorf and J. R. Rohlicek, site report to appear in *Proceedings of the ARPA Workshop on Human Language Technology,* 1993.
- "On the Use of Tied-Mixture Distributions," O. Kimball and M. Ostendorf, to appear in *Proceedings of the ARPA Workshop on Human Language Technology,* 1993.
- "Maximum Likelihood Clustering of Gaussians for Speech Recognition," A. Kannan, M. Ostendorf and J. R. Rohlicek, submitted to *IEEE Transactions on Speech and Audio Processing.*

Principal Investigator Name: Mari Ostendorf
PI Institution: Boston University
PI Phone Number: 617-353-5430
PI E-mail Address: mo@raven.bu.edu
Grant or Contract Title: Segment-Based Acoustic Models for Continuous Speech Recognition
Grant or Contract Number: ONR-N00014-92-J-1778
Reporting Period: 1 January 1993 – 31 March 1993

## 4 Transitions and DoD Interactions

This grant includes a subcontract to BBN, and the research results and software is available to them. Thus far, we have collaborated with BBN by combining the Byblos system with the SSM in N-Best sentence rescoring to obtain improved recognition performance, and we have provided BBN with papers and technical reports to facilitate sharing of algorithmic improvements. On their part, BBN has been very helpful to us in our WSJ porting efforts, providing us with WSJ data and consulting on format changes.

The recognition system that has been developed under the support of this grant and of a joint NSF-DARPA grant (NSF # IRI-8902124) is being used for automatically obtaining good quality phonetic alignments for a corpus of radio news speech under development at Boston University in collaboration with researchers at SRI International and MIT. We have requested support from the Linguistic Data Consortium to use this software to phonetically align the remainder of the corpus.

Principal Investigator Name: Mari Ostendorf
PI Institution: Boston University
PI Phone Number: 617-353-5430
PI E-mail Address: mo@raven.bu.edu
Grant or Contract Title: Segment-Based Acoustic Models for Continuous Speech Recognition
Grant or Contract Number: ONR-N00014-92-J-1778
Reporting Period: 1 January 1993 - 31 March 1993

## 5 Software and Hardware Prototypes

Our research has required the development and refinement of software systems for parameter estimation and recognition search, which are implemented in C or C++ and run on Sun Sparc workstations. No commercialization is planned at this time.

# Segment-Based Acoustic Models for Continuous Speech Recognition

*Mari Ostendorf* *J. Robin Rohlicek*

Boston University BBN Inc.
Boston, MA 02215 Cambridge, MA 02138

## PROJECT GOALS

The goal of this project is to develop improved acoustic models for speaker-independent recognition of continuous speech, together with efficient search algorithms appropriate for use with these models. The current work on acoustic modeling is focussed on: stochastic, segment-based models that capture the time correlation of a sequence of observations (feature vectors) that correspond to a phoneme; hierarchical stochastic models that capture higher level intra-utterance correlation; and multi-pass search algorithms for implementing these more complex models. This research has been jointly sponsored by DARPA and NSF under NSF grant IRI-8902124 and by DARPA and ONR under ONR grant N00014-92-J-1778.

## RECENT RESULTS

- Implemented different auditory-based signal processing algorithms and evaluated their use in recognition on the TIMIT corpus, finding no performance gains relative to cepstral parameters probably due to the non-Gaussian nature of auditory features.
- Improved the score combination technique for N-Best rescoring, through normalizing scores by sentence length to obtain more robust weights that alleviate problems associated with test set mismatch.
- Further investigated agglomerative and divisive clustering methods for estimating robust context-dependent models, and introduced a new clustering criterion based on a likelihood ratio test; obtained a slight improvement in performance with an order of magnitude reduction in storage costs.
- Extended the classification and segmentation scoring formalism to handle context-dependent models without requiring the assumption of independence of features between phone segments (using maximum entropy methods); evaluated different segmentation scores with results suggesting more work is needed in this area.
- Investigated the use of different phone sets and probabilistic multiple-pronunciation networks; no improvements were obtained on the RM corpus, though there may be gains in another domain.
- Extended the two-level segment/microsegment formalism to application in word recognition using context-dependent models; evaluated the trade-offs associated with modeling trajectories vs. (non-tied) microsegment mixtures, finding that mixtures are more useful for context-independent modeling but representation of a trajectory is more useful for context-dependent modeling.
- Investigated the use of tied mixtures at the frame level (as opposed to the microsegment level), evaluating different covariance assumptions and training conditions; developed new, faster mixture training algorithms; and achieved a 20% reduction in word error over our previous best results on the Resource Management task. Current SSM performance rates are 3.6% word error on the Oct89 test set and 7.3% word error on the Sep92 test set.

## PLANS FOR THE COMING YEAR

- Continue work in the classification and segmentation scoring paradigm; demonstrate improvements associated with novel models and/or features.
- Investigate the use of tied mixtures at the microsegment level.
- Port the BU recognition system to the Wall Street Journal (WSJ) task, 5000 word vocabulary.
- Develop a stochastic formalism for modeling intra-utterance dependencies assuming a hierarchical structure.
- Investigate unsupervised adaptation in the WSJ task domain.
- Investigate multi-pass search algorithms that use a lattice rather than N-Best representation of recognition hypotheses.

# ON THE USE OF TIED-MIXTURE DISTRIBUTIONS

*Owen Kimball, Mari Ostendorf*

Electrical, Computer and Systems Engineering
Boston University, Boston, MA 02215

## ABSTRACT

Tied-mixture (or semi-continuous) distributions are an important tool for acoustic modeling, used in many high-performance speech recognition systems today. This paper provides a survey of the work in this area, outlining the different options available for tied mixture modeling, introducing algorithms for reducing training time, and providing experimental results assessing the trade-offs for speaker-independent recognition on the Resource Management task. Additionally, we describe an extension of tied mixtures to segment-level distributions.

## 1. INTRODUCTION

Tied-mixture (or semi-continuous) distributions have rapidly become an important tool for acoustic modeling in speech recognition since their introduction by Huang and Jack [1] and Bellegarda and Nahamoo [2], finding widespread use in a number of high-performance recognition systems. Tied mixtures have a number of advantageous properties that have contributed to their success. Like discrete, "non-parametric" distributions, tied mixtures can model a wide range of distributions including those with an "irregular shape," while retaining the smoothed form characteristic of simpler parametric models. Additionally, because the component distributions of the mixtures are shared, the number of free parameters is reduced, and tied-mixtures have been found to produce robust estimates with relatively small amounts of training data. Under the general heading of tied mixtures, there are a number of possible choices of parameterization that lead to systems with different characteristics. This paper outlines these choices and provides a set of controlled experiments assessing trade-offs in speaker-independent recognition on the Resource Management corpus in the context of the stochastic segment model (SSM). In addition, we introduce new variations on training algorithms that reduce computational requirements and generalize the tied mixture formalism to include segment-level mixtures.

## 2. PREVIOUS WORK

A central problem in the statistical approach to speech recognition is finding a good model for the probability of acoustic observations conditioned on the state in hidden-Markov models (HMM), or for the case of the SSM, conditioned on a region of the model. Some of the options that have been investigated include discrete distributions based on vector quantization, as well as Gaussian, Gaussian mixture and tied-Gaussian mixture distributions. In tied-mixture modeling, distributions are modeled as a mixture of continuous densities, but unlike ordinary, non-tied mixtures, rather than estimating the component Gaussian densities separately, each mixture is constrained to share the same component densities with only the weights differing. The probability density of observation vector $\mathbf{x}$ conditioned on being in state $i$ is thus

$$p(\mathbf{x} \mid s = i) = \sum_k w_{ik} p_k(\mathbf{x}). \qquad (1)$$

Note that the component Gaussian densities, $p_k(\mathbf{x}) \sim N(\mu_k, \Sigma_k)$, are not indexed by the state, $i$. In this light, tied mixtures can be seen as a particular example of the general technique of tying to reduce the number of model parameters that must be trained [3].

"Tied mixtures" and "semi-continuous HMMs" are used in the literature to refer to HMM distributions of the form given in Equation (1). The term "semi-continuous HMMs" was coined by Huang and Jack, who first proposed their use in continuous speech recognition [1]. The "semi-continuous" terminology highlights the relationship of this method to discrete and continuous density HMMs, where the mixture component means are analogous to the vector quantization codewords of a discrete HMM and the weights to the discrete observation probabilities, but, as in continuous density HMMs, actual quantization with its attendant distortion is avoided. Bellegarda and Nahamoo independently developed the same technique which they termed "tied mixtures" [2]. For simplicity, we use only one name in this paper, and choose the term tied mixtures, to highlight the relationship to other types of mixture distributions and because our work is based on the SSM, not the HMM.

Since its introduction, a number of variants of the tied mixture model have been explored. First, different assumptions can be made about feature correlation within individual mixture components. Separate sets of tied mixtures have been used for various input features including cepstra, derivatives of cepstra, and power and its derivative, where each of these feature sets have been treated as independent observation streams. Within an observation stream, different assumptions about feature correlation have been explored, with some researchers currently favoring diagonal covariance matrices [4, 5] and others adopting full covariance matrices [6, 7].

Second, the issue of parameter initialization can be important, since the training algorithm is an iterative hill-climbing technique that guarantees convergence only to a local optimum. Many researchers initialize their systems with parameters estimated from data subsets determined by K-means clustering, e.g. [6], although Paul describes a different, bootstrapping initialization [4]. Often a large number of mixture components are used and, since the parameters can be overtrained, contradictory results are reported on the benefits of parameter re-estimation. For example, while many researchers find it useful to reestimate all parameters of the mixture models in training, BBN reports no benefit for updating means and covariances after the initialization from clustered data [7].

Another variation, embodied in the CMU senone models [8], involves tying mixture weights over classes of context-dependent models. Their approach to finding regions of mixture weight tying involves clustering discrete observation distributions and mapping these clustered distributions to the mixture weights for the associated triphone contexts.

In addition to the work described above, there are related methods that have informed the research concerning tied mixtures. First, mixture modeling does not require the use of Gaussian distributions. Good results have also been obtained using mixtures of Laplacian distributions [9, 10], and presumably other component densities would perform well too. Ney [11] has found strong similarities between radial basis functions and mixture densities using Gaussians with diagonal covariances. Recent work at BBN has explored the use of elliptical basis functions which share many properties with tied mixtures of full-covariance Gaussians [12]. Second, the positive results achieved by several researchers using non-tied mixture systems [13] raise the question of whether tied-mixtures have significant performance advantages over untied mixtures when there is adequate training data. It is possible to strike a compromise and use limited tying: for instance the context models of a phone can all use the same tied distributions (e.g. [14, 15]).

Of course, the best choice of model depends on the nature of the observation vectors and the amount of training data. In addition, it is likely that the amount of tying in a system can be adjusted across a continuum to fit the particular task and amount of training data. However, an assessment of modeling trade-offs for speaker-independent recognition is useful for providing insight into the various choices, and also because the various results in the literature are difficult to compare due to differing experimental paradigms.

## 3. TRAINING ALGORITHMS

In this section we first review properties of the SSM and then describe the training algorithm used for tied mixtures with the SSM. Next, we describe an efficient method for training context-dependent models, and lastly we describe a parallel implementation of the trainer that greatly reduces experimentation time.

### 3.1. The SSM and "Viterbi" Training with Tied Mixtures

The SSM is characterized by two components: a family of length-dependent distribution functions and a deterministic mapping function that determines the distribution for a variable-length observed segment. More specifically, in the work presented here, a linear time warping function maps each observed frame to one of $m$ regions of the segment model. Each region is described by a tied Gaussian mixture distribution, and the frames are assumed conditionally independent given the length-dependent warping. The conditional independence assumption allows robust estimation of the model's statistics and reduces the computation of determining a segment's probability, but the potential of the segment model is not fully utilized. Under this formulation, the SSM is similar to a tied-mixture HMM with a phone-length-dependent, constrained state trajectory. Thus, many of the experiments reported here translate to HMM systems.

The SSM training algorithm [16] iterates between segmentation and maximum likelihood parameter estimation, so that during the parameter estimation phase of each iteration, the segmentation of that pass gives a set of known phonetic boundaries. Additionally, for a given phonetic segmentation, the assignment of observations to regions of the model is uniquely determined. SSM training is similar to HMM "Viterbi training", in which training data is segmented using the most likely state sequence and model parameters are updated using this segmentation. Although it is possible to define an SSM training algorithm equivalent to the Baum-Welch algorithm for HMMs, the computation is prohibitive for the SSM because of the large effective state space.

The use of a constrained segmentation greatly simplifies parameter estimation in the tied mixture case, since there is only one unobserved component, the mixture mode. In this case, the parameter estimation step of the iterative segmentation/estimation algorithm involves the standard iterative expectation-maximization (EM) approach to estimating the parameters of a mixture distribution [17]. In contrast, the full EM algorithm for tied mixtures in an HMM handles both the unobserved state in the Markov chain and the unobserved mixture mode [2].

## 3.2. Tied-Mixture Context Modeling

We have investigated two methods for training context-dependent models. In the first, weights are used to combine the probability of different types of context. These weights can be chosen by hand [18] or derived automatically using a deleted-interpolation algorithm [3]. Paul evaluated both types of weighting for tied-mixture context modeling and reported no significant performance difference between the two [4]. In our experiments, we evaluated just the use of hand-picked weights.

In the second method, only models of the most detailed context (in our case triphones) are estimated directly from the data and simpler context models (left, right, and context-independent models) are computed as marginals of the triphone distributions. The computation of marginals is negligible since it involves just the summing and normalization of mixture weights at the end of training. This method reduces the number of model updates in training in proportion to the number of context types used, although the computation of observation probabilities conditioned on the mixture component densities, remains the same. In recognition with marginal models, it is still necessary to combine the different context types, and we use the same hand-picked weights as before for this purpose. We compared the two training methods and found that performance on an independent test set was essentially the same for both methods (marginal training produced 2 fewer errors on the Feb89 test set) and the marginal trainer required 20 to 35% less time, depending on the model size and machine memory.

## 3.3. Parallel Training

To reduce computation, our system prunes low probability observations, as in [4], and uses the marginal training algorithm described above. However, even with these savings, tied-mixture training involves a large computation, making experimentation potentially cumbersome. When the available computing resources consist of a network of moderately powerful workstations, as is the case at BU, we would like to make use of many machines at once to speed training. At the highest level, tied mixture training is inherently a sequential process, since each pass requires the parameter estimates from the previous pass. However, the bulk of the training computation involves estimating counts over a database, and these counts are all independent of each other. We can therefore speed training by letting machines estimate the counts ior different parts of the database in parallel and combine and normalize their results at the end of each pass.

To implement this approach we use a simple "bakery" algorithm to assign tasks: as each machine becomes free, it reads and increments the value of a counter from a common location indicating the sentences in the database it should work on next. This approach provides load balancing, allowing us to make efficient use of machines that may differ in speed. Because of the coarse grain of parallelism (one task typically consists of processing 10 sentences), we can use the relatively simple mechanism of file locking for synchronization and mutual exclusion, with no noticeable efficiency penalty. Finally, one processor is distinguished as the "master" processor and is assigned to perform the collation and normalization of counts at the end of each pass. With this approach, we obtain a speedup in training linear with the number of machines used, providing a much faster environment for experimentation.

# 4. MODELING & ESTIMATION TRADE-OFFS

Within the framework of tied Gaussian mixtures, there are a number of modeling and training variations that have been proposed. In this section, we will describe several experiments that investigate the performance implications of some of these choices.

## 4.1. Experimental Paradigm

The experiments described below were run on the Resource Management (RM) corpus using speaker-independent, gender-dependent models trained on the standard SI-109 data set. The feature vectors used as input to the system are computed at 10 millisecond intervals and consist of 14 cepstral parameters, their first differences, and differenced energy (second cepstral differences are not currently used). In recognition, the SSM uses an N-best rescoring formalism to reduce computation: the BBN BYBLOS system [7] is used to generate 20 hypotheses per sentence, which are rescored by the SSM and combined with the number of phones, number of words, and (optionally) the BBN HMM score, to rerank the hypotheses. The weights for recombination are estimated on one test set and held fixed for all other test sets. Since our previous work has indicated problems in weight estimation due to test-set mismatch, we have recently introduced a simple time normalization of the scores that effectively reduces the variability of scores due to utterance length and leads to more robust performance across test sets.

Although the weight estimation test set is strictly speaking part of the training data, we find that for most experiments, the bias in this type of testing is small enough to allow us to make comparisons between systems when both are run on the weight-training set. Accordingly some of the experiments reported below are only run on the weight training test set. Of course, final evaluation of a system must be on an independent test set.

## 4.2. Experiments

We conducted several series of experiments to explore issues associated with parameter allocation and training. The results are compared to a baseline, non-mixture SSM that uses full covariance Gaussian distributions.

The first set of experiments examined the number of component densities in the mixture, together with the choice of full- or diagonal-covariance matrices for the mixture component densities. Although the full covariance assumption provides a more detailed description of the correlation between features, diagonal covariance models require substantially less computation and it may be possible to obtain very detailed models using a larger number of diagonal models.

In initial experiments with just female speakers, we used diagonal covariance Gaussians and compared 200- versus 300-density mixture models, exploring the range typically reported by other researchers. With context-independent models, after several training passes, both systems got 6.5% word error on the Feb89 test set. For context-dependent models, the 300-density system performed substantially better, with a 2.8% error rate, compared with 4.2% for the 200 density system. These results compare favorably with the baseline SSM which has an error rate on the Feb89 female speakers of 7.7% for context-independent models and 4.8% for context-dependent models.

For male speakers, we again tried systems of 200 and 300 diagonal covariance density systems, obtaining error rates of 10.9% and 9.1% for each, respectively. Unlike the females, however, this was only slightly better than the result for the baseline SSM, which achieves 9.5%. We tried a system of 500 diagonal covariance densities, which gave only a small improvement in performance to 8.8% error. Finally, we tried using full-covariance Gaussians for the 300 component system and obtained an 8.0% error rate. The context-dependent performance for males using this configuration showed similar improvement over the non-mixture SSM, with an error rate of 3.8% for the mixture system compared with 4.7% for the baseline. Returning to the females, we found that using full-covariance densities gave the same performance as diagonal. We have adopted the use of full-covariance models for both genders for uniformity, obtaining a combined word error rate of 3.3% on the Feb89 test set. In the RM SI-109 training corpus, the training data for males is roughly 2.5 times that for females, so it is not unexpected that the optimal parameter allocation for each may differ slightly.

Unlike other reported systems which treat cepstral parameters and their derivatives as independent observation streams, the BU system models them jointly using a single output stream, which gives better performance than independent streams with a single Gaussian distribution (non-mixture system). Presumably, the result would also hold for mixtures.

Since the training is an iterative hill climbing technique, initialization can be important to avoid converging to a poor solution. In our system, we choose initial models, using one of the two methods described below. These models are used as input to several iterations of context-independent training followed by context-dependent training. We add a small padding value to the weight estimates in the early training passes to delay premature parameter convergence.

We have investigated two methods for choosing the initial models. In the first, we cluster the training data using the *K-means* algorithm and then estimate a mean and covariance from the data corresponding to each cluster. These are then used as the parameters of the component Gaussian densities of the initial mixture. In the second method, we initialize from models trained in a non-mixture version of the SSM. The initial densities are chosen as means of triphone models, with covariances chosen from the corresponding context-independent model. For each phone in our phone alphabet we iteratively choose the triphone model of that phone with the highest frequency of occurrence in training. The object of this procedure is to attempt to cover the space of phones while using robustly estimated models.

We found that the *K-means* initialized models converged slower and had significantly worse performance on independent test data than that of the second method. Although it is possible that with a larger padding valur added to the weigh^t^ estimates and more training passes, the *K-means* models might have "caught up" with the

| | Test set | |
|---|---|---|
| System | Oct 89 | Sep 92 |
| Baseline SSM | 4.8 | 8.5 |
| T.M. SSM | 3.6 | 7.3 |
| T.M. SSM + HMM | 3.2 | 6.1 |

Table 1: Word error rate on the Oct89 and Sep92 test sets for the baseline non-mixture SSM, the tied-mixture SSM alone and the SSM in combination with the BYBLOS HMM system.

other models, we did not investigate this further.

The various elements of the mixtures (means, covariances, and weights) can each be either updated in training, or assumed to have fixed values. In our experiments, we have consistently found better performance when all parameters of the models are updated.

Table 1 gives the performance on the RM Oct89 and Sept92 test set for the baseline SSM, the tied-mixture SSM system, and the tied-mixture system combined in N-best rescoring with the BBN BYBLOS HMM system. The mixture SSM's performance is comparable to results reported for many other systems on these sets. We note that it may be possible to improve SSM performance by incorporating second difference cepstral parameters as most HMM systems do.

## 5. SEGMENTAL MIXTURE MODELING

In the version of the SSM described in this paper, in which observations are assumed conditionally independent given model regions, the dependence of observations over time is modeled implicitly by the assumption of time-dependent stationary regions in combination with the constrained warping of observations to regions. Because segmentation is explicit in this model, in principle it is straightforward to model distinct segmental trajectories over time by using a mixture of such segment-level models, and thus take better advantage of the segment formalism. The probability of the complete segment of observations, $\mathbf{Y}$, given phonetic unit $\alpha$ is then

$$p(\mathbf{Y} \mid \alpha) = \sum_k w_k \, p(\mathbf{Y} \mid \alpha_k),$$

where each of the densities $p(\mathbf{Y} \mid \alpha_k)$ is an SSM. The component models could use single Gaussians instead of tied mixtures for the region dependent distributions and they would remain independent frame models, but in training all the observations for a phone would be updated jointly, so that the mixture components capture distinct trajectories of the observations across a complete segment. In practice, each such trajectory is a point in a very high-dimensional feature space, and it is necessary to reduce the parameter dimension in order to train such models. There are several ways to do this. First, we can model the trajectories within smaller, subphonetic units, as in the microsegment model described in [19, 20]. Taking this approach and assuming microsegments are independent, the probability for a segment is

$$p(\mathbf{Y} \mid \alpha) = \prod_j \sum_k w_{jk} \, p(\mathbf{Y}_j \mid \alpha_{jk}), \tag{2}$$

where $\alpha_{jk}$ is the $k^{th}$ mixture component of microsegment $j$ and $\mathbf{Y}_j$ is the subset of frames in $\mathbf{Y}$ that map to microsegment $j$. Given the SSM's deterministic warping and assuming the same number of distributions for all mixture components of a given microsegment, the extension of the EM algorithm for training mixtures of this type is straightforward. The tied-mixture SSM discussed in previous sections is a special case of this model, in which we restrict each microsegment to have just one stationary region and a corresponding mixture distribution.

A different way to reduce the parameter dimension is to continue to model the complete trajectory across a segment, but assume independence between subsets of the features of a frame. This case can be expressed in the general form of (2) if we reinterpret the $\mathbf{Y}_j$ as vectors with the same number of frames as the complete segment, but for each frame, only a specific subset of the original frame's features are used. We can of course combine these two approaches, and assume independence between observations representing feature subsets of different microsegmental units. There are clearly a large number of possible decompositions of the complete segment into time and feature subsets, and the corresponding models for each may have different properties. In general, because of constraints of model dimensionality and finite training data, we expect a trade-off between the ability to model trajectories across time and to model the correlation of features within a local time region.

Although no single model of this form may have all the properties we desire, we do not necessarily have to choose one to the exclusion of all others. All the models discussed here compute probabilities over the same observation space, allowing for a straightforward combination of different models, once again using the simple mechanism of non-tied mixtures:

$$p(\mathbf{Y} \mid \alpha) = \sum_i \prod_j \sum_k w_{ijk} \, p(\mathbf{Y}_j \mid \alpha_{ijk}).$$

In this case, each of the $i$ components of the leftmost summation is some particular realization of the general model expressed in Equation (2). Such a mixture can combine component models that individually have beneficial properties for modeling either time or frequency correlation, and the combined model may be able to model both aspects well. We note that, in principle, this model can also be extended to larger units, such as syllables or words.

## 6. SUMMARY

This paper provided an overview of work using tied-mixture models for speech recognition. We described the use of tied mixtures in the SSM as well as several innovations in the training algorithm. Experiments comparing performance for different parameter allocation choices using tied-mixtures were presented. The performance of the best tied-mixture SSM is comparable to HMM systems that use similar input features. Finally, we presented a general method we are investigating for modeling segmental dependence with the SSM.

## ACKNOWLEDGMENTS

The authors gratefully acknowledge BBN Inc. for their help in providing the N-best sentence hypotheses. We thank J. Robin Rohlicek of BBN for many useful discussions. This research was jointly funded by NSF and DARPA under NSF grant number IRI-8902124, and by DARPA and ONR under ONR grant number N00014-92-J-1778.

## References

1. Huang, X. D. and Jack, M. A., "Performance comparison between semi-continuous and discrete hidden Markov models," *IEE Electronics Letters*, Vol. 24 no. 3, pp. 149-150.
2. Bellegarda, J. R. and Nahamoo, D., "Tied Mixture Continuous Parameter Modeling for Speech Recognition," *IEEE Trans. on Acoust., Speech and Signal Processing*, Dec 1990, pp. 2033–2045.
3. Jelinek, F. and Mercer, R.L., "Interpolated Estimation of Markov Source Parameters from Sparse Data," in *Proc. Workshop Pattern Recognition in Practice*, May 1980, pp. 381-397.
4. Paul, D.B., "The Lincoln Tied-Mixture HMM Continuous Speech Recognizer," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, May 1991, pp. 329–332.
5. Murveit, H., Butzberger, J., Weintraub, M., "Speech Recognition in SRI's Resource Management and ATIS Systems," *Proc. of the DARPA Workshop on Speech and Natural Language*, June 1990, pp. 94–100.
6. Huang, X.D., Lee, K.F., Hon, H.W., and Hwang, M.-Y., "Improved Acoustic Modeling with the SPHINX Speech Recognition System," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, May 1991, pp. 345–348.
7. Kubala, F., Austin, S., Barry, C., Makhoul, J. Placeway, P., and Schwartz, R., "BYBLOS Speech Recognition Benchmark Results," *Proc. of the DARPA Workshop on Speech and Natural Language*, Asilomar, CA, Feb. 1991, pp. 77-82.
8. Hwang, M.-Y., Huang, X. D., "Subphonetic Modeling with Markov States - Senone," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, March 1992, pp. I-33–36.
9. Ney, H., Haeb-Umbach, R., Tran, B.-H., Oerder, M., "Improvements in Beam Search for 10000-Word Continuous Speech Recognition," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, April 1992, pp. I-9-12.
10. Baker, J. K., Baker, J. M.. Bamberg, P., Bishop, K., Gillick, L., Helman, V., Huang, Z., Ito, Y., Lowe, S., Peskin, B., Roth, R., Scattone, F., "Large Vocabulary Recognition of Wall Street Journal Sentences at Dragon Systems," *Proc. of the DARPA Workshop on Speech and Natural Language*, February 1992, pp. 387-392.
11. H. Ney, "Speech Recognition in a Neural Network Framework: Discriminative Training of Gaussian Models and Mixture Densities as Radial Basis Functions," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, May 1991, pp. 573–576.
12. Zavaliagkos, G., Zhao, Y., Schwartz, R., and Makhoul,J., to appear in *Proc. of the DARPA Workshop on Artificial Neural Networks and CSR*, Sept. 1992.
13. Pallett, D., Results for the Sept. 1992 Resource Management Benchmark, presented at the DARPA Workshop on Artificial Neural Networks and CSR, Sept. 1992.
14. Lee, C., Rabiner, L., Pieraccini, R., and Wilpon, J., "Acoustic Modeling for Large Vocabulary Speech Recognition," *Computer Speech and Language*, April. 1990, pp. 127–165.
15. Paul, D. B., "The Lincoln Robust Continuous Speech Recognizer," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, May 1989, pp. 449–452.
16. Ostendorf, M. and Roukos, S. , "A Stochastic Segment Model for Phoneme-Based Continuous Speech Recognition," *IEEE Trans. on Acoust., Speech and Signal Processing*, Dec. 1989, pp. 1857–1869.
17. Dempster, A., Laird, N. and Rubin, D., "Maximum Likelihood from Incomplete Data via the EM Algorithm," *J. Royal Statist. Soc. Ser. B*, Vol. 39 No. 1, pp. 1–22, 1977.
18. Schwartz, R., Chow, Y. L., Kimball, O., Roucos, S., Krasner, M. and Makhoul, J., "Context-Dependent Modeling for Acoustic-Phonetic Recognition of Continuous Speech," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, March 1985, pp. 1205–1208.
19. Digalakis, V. *Segment-Based Stochastic Models of Spectral Dynamics for Continuous Speech Recognition*, Boston University Ph.D. Dissertation, 1992.
20. Kannan, A., and Ostendorf, M., "A Comparison of Trajectory and Mixture Modeling in Segment-Based Word Recognition," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, April 1993.

# Maximum Likelihood Clustering of Gaussians for Speech Recognition*

A. Kannan† M. Ostendorf† J. R. Rohlicek‡

† Boston University ‡ BBN Inc.

Contact: Mari Ostendorf

ECS Department, Boston University, 44 Cummington Street, Boston, MA 02215

Phone: (617) 353-5430 Fax: (617) 353-6440 Email: mo@raven.bu.edu

April 3, 1993

EDICS SA 1.6.4

**Abstract**

This correspondence describes a method for clustering multivariate Gaussian distributions using a Maximum Likelihood criterion. We point out possible applications of model clustering, and then use the approach to determine classes of shared covariances for context modeling in speech recognition, achieving an order of magnitude reduction in the number of covariance parameters, with no loss in recognition performance.

## 1 Introduction

Distribution clustering is an important tool in statistical modeling. In speech recognition in particular, distribution clustering can be used to reduce the number of context-dependent models (which enables robust parameter estimates and reduces recognition computation and storage costs), to provide an initial estimate of component distributions for mixture models, and to group similar models for building an initial "fast-match" function in large vocabulary recognition. This work describes a new method

*This research was jointly funded by NSF and DARPA under NSF grant number IRI-8902124, and by DARPA and ONR under ONR grant number N00014-92-J-1778.

of distribution clustering that handles continuous observations and is consistent with a Maximum Likelihood (ML) parameter estimation criterion.

Two important issues associated with clustering distributions include the general method (agglomerative or divisive hierarchical methods vs. K-means clustering) and the clustering criterion or objective function. Initial work in clustering models for speech recognition used variations on agglomerative clustering [1, 2]. Subsequent work explored divisive clustering methods, using linguistically motivated questions (partitioning functions) about phonetic context for splitting the data [3, 4]. An advantage of this divisive clustering approach, as Lee *et al.* point out [3], is that conditioning contexts unseen in training can be easily mapped to a cluster that provides a robust but detailed model. For this reason, our work uses divisive clustering, although the similarity criterion we propose could easily be applied to agglomerative clustering as well.

The second issue in clustering is the choice of a clustering criterion or objective function. One possibility is to use a measure of distribution similarity, such as information divergence (see [5] for the hidden Markov model (HMM)) or the chi-squared-like measure used in [1] for Gaussian distributions. However, such similarity measures tend to be more useful for agglomerative clustering than for divisive clustering, because agglomerative clustering does not require the computation of a centroid associated with the similarity measure and the centroid is difficult to define for these criteria. In addition, similarity measures on distributions may not faithfully represent the similarity of the data from which the distributions were estimated, particularly if distribution assumptions were inaccurate or parameters were estimated from sparse data. Other objective functions proposed for distribution clustering include entropy measures [3] and likelihood ratios ([4] for discrete observations and [6] for Gaussian distributions). The likelihood ratio criterion represents the relative probability of a set of data using one vs. two models, and its use in divisive clustering guarantees an increase in the likelihood of the data. Thus the likelihood ratio criterion has the advantage that it is consistent with the objective of maximum likelihood parameter estimation, that is if the clustered distributions and not just the cluster definition are used in the model (the distinction between our use of clustering in estimating the model parameters and the use of clustering in [4] to determine regions of parameter tying).

This work investigates the use of a likelihood ratio criterion in divisive clustering for context-dependent modeling in speech recognition, extending the work of Gish *et al.* [6] which looked at agglomerative clustering for speaker segmentation and identification. We describe general methods for clustering data to determine appropriate multivariate Gaussian models under different parameter tying conditions, and then present experiments in clustering covariances, specifically for estimating Gaussian distributions that represent regions of a phoneme segment as used in the Stochastic Segment Model (SSM). (The region-dependent distributions in the SSM are analogous to state-dependent observation distributions in an HMM.) Note that for the speech recognition application, the question of whether to cluster on the phone level or sub-phone level arises. Like [7] but unlike most other reported work, the experiments here focus on the sub-phone level, though the general method is also applicable to phone-level clustering. The results show that the number of covariance parameters can be reduced by more than a factor of ten with clustering, with no loss in recognition performance.

## 2 Clustering Paradigm

The clustering algorithm is a binary tree growing procedure, similar to decision tree design [8], that successively partitions the observations (splits a node in the tree), at each step minimizing a splitting criterion over a pre-determined set of allowable binary partitions. For each allowable binary partition of the data, we evaluate a likelihood ratio to choose between one of two hypotheses:

- $H_0$: the observations were generated from one distribution (that corresponds to the ML estimate for the parent node).
- $H_1$: the observations were generated from two different distributions (that correspond to the ML estimates for the child nodes), and

The likelihood ratio, $\lambda$, is defined as the ratio of the likelihood of the observations being generated from one distribution ($H_0$) to the likelihood of the observations in the partition being generated from two different distributions ($H_1$). For Gaussians, $\lambda$ can be expressed as a product of the quantities $\lambda_{COV}$ and $\lambda_{MEAN}$ [6], which are expressed in terms of the sufficient statistics of the observation sets:

$$\lambda_{MEAN} = \left(1 + \frac{n_l n_r}{n^2}(\hat{\mu}_l - \hat{\mu}_r)^t W^{-1}(\hat{\mu}_l - \hat{\mu}_r)\right)^{\frac{-n}{2}} \tag{1}$$

$$\lambda_{COV} = \left(\frac{|\hat{\Sigma}_l|^{\alpha}|\hat{\Sigma}_r|^{(1-\alpha)}}{|W|}\right)^{\frac{n}{2}} \tag{2}$$

where $n_l$ and $n_r$ are the number of observations in the left and right child nodes with $n = n_l + n_r$, $\hat{\mu}_l$ and $\hat{\mu}_r$ are the sample means of the left and right nodes, $\hat{\Sigma}_l$ and $\hat{\Sigma}_r$ are the sample covariances associated with the left and right nodes, $\alpha = \frac{n_l}{n}$, and $W$ is the frequency weighted tied covariance, viz., $W = \frac{n_l}{n}\hat{\Sigma}_l + \frac{n_r}{n}\hat{\Sigma}_r$.

There are different variations on clustering with the likelihood ratio criterion, corresponding to different hypothesis tests on the candidate partition of a node. If the clustering is to determine whether the observations in two sets share a common covariance only with unspecified means, then the increase in log likelihood is given by $-\log \lambda_{COV}$. Alternatively, if the hypothesis test is over the complete distributions, then the increase in likelihood due to the partition is $-(\log \lambda_{COV} + \log \lambda_{MEAN})$. Finally, if the distributions are assumed to share a common covariance and only distribution means are to be clustered, then the likelihood ratio criterion is $-\log \lambda_{MEAN}$. (Note that the mean clustering case would require a hybrid divisive plus K-means clustering to guarantee increase in likelihood, since the common covariance is defined as the sample covariance of the parent node.) The derivations for these different cases can be found in [9].

Divisive clustering involves growing a binary tree using a greedy algorithm for maximizing the likelihood of the data. For each terminal node in the tree, we evaluate the increase in likelihood for all binary partitions allowed, and then split the terminal node with the partition that results in the largest increase in likelihood. The tree is grown until there are no more splits that result in valid child nodes. Here, it is assumed that valid terminal nodes must have more than $T_c$ observations, where $T_c$ is an empirically determined threshold to indicate that a reliable covariance can be estimated for that node (we use $T_c = 250$, for vector dimension 29). The full tree can be used for the set of clustered models as in the experiments described here, or alternatively, one could use tree pruning techniques [10] to determine the appropriate number of distributions.

This technique to cluster Gaussians can be used for clustering context-dependent acoustic models in speech recognition. In this work, we cluster triphones, where a triphone is a phone conditioned on the phone label of its left and right neighbor, but the conditioning contexts could potentially include a larger window of neighbors [4] or information such as lexical stress. More specifically, divisive clustering is performed independently on the observations that correspond to each region of the center phone, with the goal of finding classes of triphones that can share a common covariance. The partitions used to test the likelihood ratio are found by asking linguistically motivated questions related to features such as the place and manner of articulation of the immediate left and right neighboring phones of the triphone. Only simple questions (i.e., questions about one variable) are used in this implementation; a method for designing trees with compound questions is described in [3]. The triphone clustering framework is illustrated in Figure 1.

When the tree is grown, each terminal node has a set of observations associated with it that map to a set of triphone distributions. The partition of observations directly implies a partition of triphones, since the allowable questions refer to the left and right neighboring phone labels. Each node is associated with a covariance, which is an unbiased estimate of the tied covariance for the constituent distributions computed by taking a weighted average of the separate triphone-dependent covariances. During recognition, all distributions associated with a terminal node share this covariance. Although it would have been possible to cluster means as well, we simply used the triphone-dependent means and backed off to combined left- and right-context-dependent means when necessary due to insufficient triphone training data.

# 3 Experiments

We conducted experiments to assess the proposed method of clustering triphones in continuous speech recognition. In this section, we describe the recognition paradigm and then present results.

## 3.1 Paradigm

We evaluated the effects of clustering triphones for the Stochastic Model for representing variable-duration phonemes, first introduced in [11]. In brief, the SSM assumes that each segment generates an observation sequence $Y = [y_1, \ldots, y_L]$ of random length $L$ using a model for each phone $\alpha$ consisting of 1) a family of joint density functions (one for every observation length), and 2) a collection of mappings that specify the particular density function for a given observation length. Typically, the model assumes that segments are described by a fixed-length sequence of locally time-invariant regions (or regions of tied distribution parameters). A deterministic mapping specifies which region corresponds to each observation vector.

The specific version used here [12] assumes that frames within a segment are conditionally independent given the segment length. In this case, the probability of a segment given phone $\alpha$ is the product of the probability of each observation $y_i$ and the probability of its (known) duration $L$:

$$p(Y|\alpha) = p(Y, L|\alpha) = p(L|\alpha) \prod_{i=1}^{L} p(y_i|\alpha, T_L(i)),$$

where the distribution used corresponds to region $T_L(i)$. The distributions associated with a region $j$, $P(y|\alpha, j)$, are multivariate Gaussians. The phone length distribution $p(L|\alpha)$ is a smoothed relative frequency estimate in this work. $T_L(i)$ determines the mapping of the $L$-long observation to the $m$ regions in the model. The function $T_L(i)$ in this work is linear in time for the entire segment.

To reduce the computational costs associated with a segment-based model, which has a much higher effective search space than an HMM, we use the N-best rescoring formalism for continuous word recognition [13]. In this formalism, one recognition system produces the top N hypotheses for an utterance, the hypotheses rescored by other knowledge sources, and the different scores are combined to rerank the hypotheses. In addition to reducing computation for the SSM (by reducing the search space), the N-best rescoring paradigm provides a mechanism for integrating very different types of knowledge sources, though this aspect is not explored here. For these experiments, the initial list of candidate sentences were generated using BBN's BYBLOS system and then rescored by the SSM. The BYBLOS system is an HMM-based system that uses tied Gaussian mixtures and context-dependent models including cross-word triphones [14]. Once the N-best list is rescored by the SSM, it is reordered according to a linear combination of the SSM log acoustic score, the number of words in the sentence (insertion penalty) and the number of phonemes in the sentence. We estimate the set of weights in the linear combination that minimizes average word error in the top ranking hypotheses [15].

Results are reported on the speaker-independent Resource Management task (continuous speech, 991 word vocabulary). The SSM models are trained on the SI-109, 3990 utterance SI training set. The training was partitioned to obtain gender-dependent models; the specific gender used by the SSM in recognition was determined by the BBN system for detecting gender. The BU SSM system uses frame-based observations of spectral features, including 14 mel-warped cepstra and their first differences, plus the first difference of log energy. The segment model uses a sequence of $m = 8$ multivariate (full) Gaussian distributions, assuming frames are conditionally independent given the segment length. In these experiments, we use N = 20 for the N-best list. The correct sentence is included in this list about 98% of the time by the Byblos system, using the word-pair grammar.

### 3.2 Results

The February 89 speaker-independent (SI) test set was used to estimate gender-independent weights that were then used to combine scores for the evaluation test set (October 89). Recognition performance was computed as the word error raıe based on the top ranking hypotheses after rescoring. The performance of our system on the October 89 test set was 5.0% for the tied covariance (base-line) system, 4.9% for the system using clustering with the full likelihood criterion and 4.9% when clustering with only the $\lambda_{COV}$ likelihood criterion. The corresponding numbers for the February 89 development test set were 4.6%, 4.2% and 4.1%, respectively. Although the performance differences on the October 89 test set are not significant, the performa..ce on the development test set provides some evidence that clustering with the theoretically appropriate $\lambda_{COV}$ criterion is also a good choice in terms of recognition performance. The results are consistent with those reported by others, in that the main benefit of clustering is a reduction in model complexity rather than an improvement in performance. In these experiments, we reduced the number of covariance parameters required by more than a factor of ten with no loss in recognition performance, and further reduction may be possible. Since the covariance parameters are the dominating factor in computation and storage costs, this represents a significant overall reduction.

## 4 Conclusions

In summary, we have described a divisive clustering paradigm for multivariate Gaussians based on a likelihood ratio test. In the context of speech recognition, we use the clustering formalism to determine classes of triphones over which to tie covariances in the SSM, finding that we can reduce the number of covariances by more than a factor of ten without any loss in recognition performance. This method will be useful for any pattern recognition problem where features are modeled using Gaussian distributions, including HMMs. This approach to clustering may also be useful for providing initial estimates of components in tied-mixtures, or determining classes of like models for designing fast initial search procedures in large vocabulary recognition.

## References

[1] D. B. Paul and E. A. Martin, "Speaker Stress-resistant Continuous Speech Recognition," *Proc. of the Inter. Conf. on Acoust., Speech and Signa' Proc.*, pp. 283-286, April 1988.

[2] K.-F. Lee, "Context-Dependent Phonetic Hidden Markov Models for Speaker-Independent Continuous Speech Recognition," *IEEE Trans. on Acoust., Speech, and Signal Proc.*, pp. 599-609, April 1990.

[3] K.-F. Lee, S. Hayamizu, H.-W. Hon, C. Huang, J. Swartz and R. Weide, "Allophone Clustering for Continuous Speech Recognition," *Proc. of the Inter. Conf. on Acoust., Speech and Signal Proc.*, pp. 749-752, April 1990.

[4] L. R. Bahl, P. V. de Souza, P. S. Gopalakrishnan, D. Nahamoo and M. A. Picheny, "Context Dependent Modeling of Phones in Continuous Speech Using Decision Trees," *Proc. DARPA Speech and Natural Language Workshop*, pp. 264-269, February 1991.

[5] B.-H. Juang and L. R. Rabiner, "A Probabilistic Distance Measure for Hidden Markov Models," *AT&T Technical Journal*, Vol. 64, No. 2, pp. 391-408, 1985.

[6] H. Gish, M. Siu, R. Rohlicek, "Segregation of Speakers for Speech Recognition and Speaker Identification", *Proc. of the Inter. Conf. on Acoust., Speech and Signal Proc.*, pp. 873-876, May 1991.

[7] M.-Y. Hwang and X. Huang, "Subphonetic Modeling for Speech Recognition," *Proc. DARPA Speech and Natural Language Workshop*, pp. 174-179, February 1992.

[8] L. Breiman, J. Friedman, R. Olshen and C. Stone, *Classification and Regression Trees*, Wadsworth and Brooks/Cole Advanced Books and Software, Montery, CA, 1984.

[9] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*, J. Wiley & Sons, New York, pp. 404-450, 1984.

[10] P. Chou, T. Lookabaugh and R. Gray, "Optimal Pruning with Applications to Tree-Structured Source Coding and Modeling," *IEEE Trans. on Information Theory*, pp. 299-315, March 1989.

[11] M. Ostendorf and S. Roukos, "A Stochastic Segment Model for Phoneme-Based Continuous Speech Recognition," *IEEE Trans. on Acoust., Speech, and Signal Proc.*, pp. 1857-1869, December 1989.

[12] M. Ostendorf, A. Kannan, O. Kimball and J. R. Rohlicek, "Continuous Word Recognition Based on the Stochastic Segment Model," *Proceedings of the 1992 DARPA Workshop on Artificial Neural Networks and Continuous Speech Recognition*, to appear.

[13] M. Ostendorf, A. Kannan, S. Austin, O. Kimball, R. Schwartz, J. R. Rohlicek, "Integration of Diverse Recognition Methodologies Through Reevaluation of N-Best Sentence Hypotheses," *Proc. DARPA Speech and Natural Language Workshop*, pp. 83-87, February 1991.

[14] F. Kubala, S. Austin, C. Barry, J. Makhoul, P. Placeway, R. Schwartz, "BYBLOS Speech Recognition Benchmark Results," *Proceedings of the DARPA Workshop on Speech and Natural Language*, pp. 77-82, February 1991.

[15] A. Kannan, M. Ostendorf, J. R. Rohlicek, "Weight Estimation for N-Best Rescoring," *Proc. DARPA Speech and Natural Language Workshop*, pp. 455-456, February 1992.

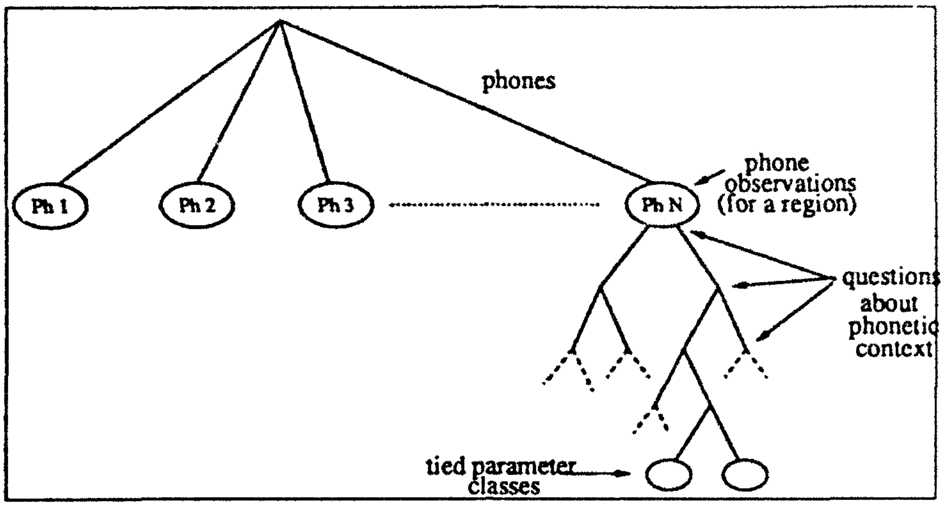
phones
Ph 1
Ph 2
Ph 3
Ph N
phone observations (for a region)
questions about phonetic context
tied parameter classes

Figure 1: Illustration of divisive clustering.